Assessment of additional phase energy losses caused by phase imbalance for data-scarce LV networks

Lurui Fang [1], Kang Ma[*] [1]

[1] Department of Electronic and Electrical Engineering, University of Bath, Bath, UK
[*]K.Ma@Bath.ac.uk

Abstract: Unbalanced phase currents, which flow in transformer windings and distribution wires, cause a significant increase (approximately 33%) of phase energy losses in low voltage (LV, 415V) networks. However, these additional phase energy losses (APEL) are hard to calculate for most LV networks. A key challenge is that these LV networks are data-scarce, with only yearly average and maximum phase currents. To estimate the APEL for data-scarce LV networks, this paper proposes a statistical approach that effectively overcomes the above challenge. Firstly, the approach calculates APEL for a sample set of data-rich networks with year-round time-series phase current data. Secondly, features are extracted from these networks by considering: 1) whether the features are strongly correlated to additional phase energy losses; and 2) whether the features can be derived from available data (e.g. yearly average and maximum phase currents) from data-scarce networks. Thirdly, to approximate mappings from the features (derived in stage 2) to the APEL (derived in stage 1), a kernel-based regression model is developed, using the above customised features. Given any data-scarce network, its APEL is then estimated by applying the regression model. Cross-validation shows that the statistical approach incurs an average error of 13% for 90% of the data-scarce LV networks, excluding the networks with very low APEL values.

1. Introduction

Phase imbalance is a widespread problem in low voltage (415V, LV) networks in the UK and other countries [1], [2], [3]. According to the data from Western Power Distribution (WPD, a UK distribution network operator), more than 50% of the LV networks suffer from a notable degree of phase imbalance. It is common that the current on the heaviest phase is greater than that on the lightest phase by more than 50% [4]. It should be noted that even if a network has perfectly balanced three phases, there is still an I^2R loss on the phase conductors because of conductor impedance. However, if the three phases are unbalanced, the I^2R loss on the phase conductors would be greater than if the three phases were balanced. The difference is the additional phase energy loss (APEL).

These unbalanced phase currents cause a significant increase in energy losses on the three phases of LV networks: 1) on distribution lines, APELs account for up to 33% of wire energy losses [4]; and 2) in distribution transformers, APELs account for up to 27% energy losses of transformer copper losses [4]. However, APELs are hard to calculate for most LV networks. A key challenge arises from the APEL estimation: a lack of time-series phase current data for the majority of LV networks that are data-scarce. These networks only have yearly average and maximum phase current data, collected once a year.

One solution to address the data scarcity challenge is to deploy monitoring devices for more than 900,000 LV networks in the UK. However, this causes a substantial cost. With sufficient data collected, a number of references assess energy losses caused by phase imbalance. Reference [5] assesses the additional copper losses caused by imbalanced loading for LV transformers. Reference [6] evaluates energy losses in distribution networks with imbalanced three phases. The APELs are calculated for networks with full data. Reference [7] develops network component models (includes load, line, and transformers) to calculate the energy losses for distribution transformers and lines. The APELs are then derived from this model. Reference [2] develops new phase imbalance indices, which are then used to estimate energy losses. Reference [8] uses complex unbalance factor to evaluate the APEL.

Reference [9] develops a combination of clustering and classification approach to estimate the imbalance-induced energy losses for data-scarce networks. However, Reference [9] focuses on the energy losses in the neutral and ground, caused by phase residual currents. Such energy losses have a fundamentally different mechanism from the APEL, which occurs on the phases. Because of the different mechanisms, this paper uses a completely different methodology from that in [9]. Compared to Reference [9] which uses a combination of clustering and classification, this paper develops a straightforward regression model using customised features. This model achieves a greater estimation accuracy than the approach in Reference [9].

In addition, it is popular to use the load loss factor to estimate the energy losses on each phase [10]. The APEL can be directly calculated if the energy losses on each of the three phases were available. However, the load loss factor k is suggested to be updated every month [10]. This incurs a prohibitively high cost to collect these data every month for the mass population of LV networks throughout the UK. Reference [11] models the correlation between the increase of energy losses and imbalance degrees based on three scenarios, e.g. 1) one phase is overloaded and the other two phases have light loads; 2) two phases are overloaded and the other phase has a light load; and 3) the three phases are overloaded, moderately loaded, and lightly loaded, respectively. Reference [12] develops a statistical approach to estimate energy losses in distribution components (e.g. distribution lines, transformer, etc.) based on load curves. However, Reference [12] does not assess the APEL caused by phase imbalance.

Based on the literature review, a research question arises: to assess the APEL caused by phase imbalance for data-scarce LV networks. This paper makes an original contribution by answering the above research question for the

first time. To this end, this paper develops a new customised statistical approach, using customised features, to assess the APELs for data-scarce networks. This approach learns the knowledge from a sample set of 800 data-rich networks (with time-series phase current data throughout a year), then infers the APELs by extrapolating the knowledge to these data-scarce networks.

The customised methodology is designed to be highly practical for distribution network operators (DNOs), who can directly apply the methodology to their business areas. Furthermore, the APEL is one of the key inputs for the cost-benefit analysis of phase rebalancing for data-scarce networks. In addition, it can help the DNOs to assess the additional heating caused by phase imbalance for data-scarce LV networks. This additional heating is one of the key components in analysing the thermal ratings of electric apparatuses (e.g. distribution transformers and lines) in data-scarce LV networks.

The rest of this paper is organized as follows: Section 2 presents the methodology. Section 3 performs case studies. Section 4 concludes this paper.

2. Methodology

The statistical approach consists of three stages. Firstly, it calculates the APELs for 800 data-rich networks with time-series phase current data throughout a year. Then, features are selected by considering: 1) whether the features are strongly correlated to the APELs; and 2) whether the features can be obtained from data-scarce networks that only have yearly average and maximum phase currents. Thirdly, a regression model is developed to map the features (derived in Stage 2) to the APELs (derived in Stage 1). Given any data-scarce network that has the feature vector as the input, the APEL is estimated by applying the developed regression model.

The flowchart of the proposed approach is shown as follows:

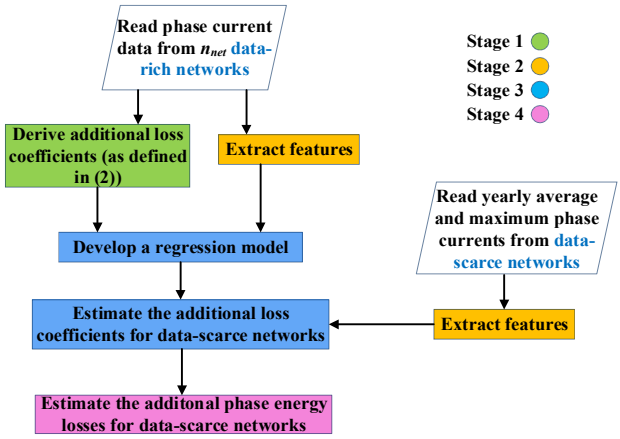

Fig. 1 *Overview of the statistical approach*

The project "Low Voltage Network Template" provides time-series phase current data throughout a year from n_{net} ($n_{net} = 800$) data-rich networks. These networks cover: 1) a good mixture of urban, suburban and rural areas; and 2) a good mixture of household, commercial and industry loads [4].

2.1. Data processing

For data-rich networks, a virtual current is defined as:

$$I_v(t) = \sqrt{I_a(t)^2 + I_b(t)^2 + I_c(t)^2 - 3\left(\frac{I_a(t) + I_b(t) + I_c(t)}{3}\right)^2} \quad (1)$$

where $I_a(t), I_b(t)$, and $I_c(t)$ denote the currents on phase a, b and c, respectively, at time t.

Then an additional loss coefficient is defined as:

$$L_{ac} = \frac{1}{n_y} \sum_{t=1}^{n_y} I_v(t)^2 \quad (2)$$

where I_v is defined in (1); n_y is the length of time-series phase current data throughout a year. The reason for defining this coefficient is to normalise the sum of $I_v(t)^2$ for all data-rich networks. This prevents large values of the sums of $I_v(t)^2$ from causing large root-mean-squared errors, thus improving the accuracy of the regression model.

For most LV networks, their topologies are unknown for the DNO. According to reference [13], loads are assumed to be distributed in a rectangular fashion along the LV networks. This results in the equivalent distribution line resistance being discounted to only 1/3 of the original line resistance, but the transformer resistance is unaffected. Therefore, the APEL is given by [13]:

$$E_{al} = T \cdot L_{ac} \cdot \left(\tfrac{1}{3} R_D + R_T\right) \quad (3)$$

where T ($T = 8760$) is the number of hours throughout a year; R_D is the resistance of the distribution line; R_T is the resistance of the transformer winding referred to the LV side.

The resistance values of distribution lines and transformers vary in different LV networks. The key output of this stage is the additional loss coefficient L_{ac}, which will be used for regression later.

2.2. Feature extraction

To select the features, two factors are considered: 1) whether the features are strongly correlated to additional loss coefficients (derived in Section 2 – 2.1); and 2) whether the features can be derived from the available data (i.e. yearly average and maximum phase currents) from data-scarce networks. Based on the above principles, four features are selected: hypothetical virtual current, maximum current, hypothetical degree of phase imbalance, and root-mean-square of unbalance ratio.

1) The hypothetical virtual current is given by:

$$I_{hv} = \sqrt{{I_{ya}}^2 + {I_{yb}}^2 + {I_{yc}}^2 - 3\left(\frac{I_{ya} + I_{yb} + I_{yc}}{3}\right)^2} \quad (4)$$

where I_{ya}, I_{yb}, I_{yc} denotes the yearly average phase currents on phases a, b and c, respectively.

2) The maximum current is given by:

$$I_m = \max\{I_{yma}, I_{ymb}, I_{ymc}\} \quad (5)$$

where I_{yma}, I_{ymb} and I_{ymc} denote the yearly maximum currents on phases a, b and c, respectively; $\max\{...\}$ indicates the maximum value of $\{...\}$.

3) The hypothetical degree of phase imbalance is given by:

$$DIB_v = \frac{(\max\{I_{ya}, I_{yb}, I_{yc}\} - \frac{I_{ya} + I_{yb} + I_{yc}}{3})}{I_{ya} + I_{yb} + I_{yc}} \quad (6)$$

where I_{ya}, I_{yb} and I_{yc} are defined in (4).

4) The root-mean-square of unbalance ratio (*RMS*)

Before deriving this *RMS* value, the positive, negative and zero sequence currents are given by:

$$\begin{bmatrix} \dot{I}_1 \\ \dot{I}_2 \\ \dot{I}_0 \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 1 & q & q^2 \\ 1 & q^2 & q \\ 1 & 1 & 1 \end{bmatrix}\begin{bmatrix} \dot{I_{ya}} \\ \dot{I_{yb}} \\ \dot{I_{yc}} \end{bmatrix} \quad (7)$$

where q is $e^{j2\pi/3}$; $\dot{I_{ya}}, \dot{I_{yb}}, \dot{I_{yc}}$ are the yearly average complex current values on phases a, b and c, respectively; the upper dot indicates that these values are complex values, which are 120 degrees apart from each other.

RMS is then given by [14]:

$$RMS = \sqrt{\left|\dot{I}_0\right|^2 + \left|\dot{I}_2\right|^2} \Big/ \left|\dot{I}_1\right| \quad (8)$$

where $\left|\dot{I}_1\right|$, $\left|\dot{I}_2\right|$ and $\left|\dot{I}_0\right|$ are the magnitudes of $\dot{I}_1, \dot{I}_2$ and $\dot{I}_0$, respectively. A feature vector consisting of the above features is given by:

$$f_v = [I_{hv}, I_{hm}, DIB_h, RMS] \quad (9)$$

where I_{hv}, I_{hm}, DIB_h and RMS are defined in (4), (5), (6), and (8), respectively.

Through the case study, a high regression accuracy is achieved when considering all the above features. This shows a strong correlation between the selected features and the additional loss coefficients.

2.3. Develop the regression model

In this stage, a kernel-based robust linear regression model is developed. It approximates the mappings from the features (derived in Section 2.2) to the additional loss coefficients (derived in Section 2.1) through training on the sample set of the data-rich networks. Then the developed mapping is applied to any data-scarce LV network with the feature vector only to estimate its additional loss coefficient. This value is then converted to the APEL for the data-scarce LV network by applying (3). The reasons for using the kernel-based robust linear regression model are: 1) robust linear regression a classic regression method [15]; 2) it is less sensitive to outliers [15]; and 3) the method allows for a higher regression accuracy compared to alternative classical regression methods. The comparison will be demonstrated in case studies.

In the first step, a quadratic kernel transformation is used to transform the feature vector from its original space to a vector in a high dimensional Hilbert space [16]. This is because the mapping in the original space is non-linear; the quadratic kernel transformation enables a nearly linear mapping in the high dimensional space. Through such a transformation, the regression accuracy is improved by 43% compared to the ordinary robust linear regression. The quadratic kernel transformation is given by:

$$f_{kv} = [k(f_{v,1}, f_{v,1}), \cdots k(f_{v,i}, f_{v,j}) \cdots, k(f_{v,4}, f_{v,4})] \quad (10)$$

where $k(f_{v,i}, f_{v,j}) = ({f_{v,i}}^{\mathrm{T}} \cdot f_{v,j} + c)^2$

where $f_{v,i}$ and $f_{v,j}$ are the i_{th} and j_{th} variables in the feature vector f_v (as defined in (9)), respectively; c denotes a constant value. Based on this transformation, the feature vector f_v is transformed into a high dimensional kernel feature vector f_{kv}. In this study, f_{kv} is a vector with 16 variables.

Then, a robust linear regression model is developed to approximate the mapping from the kernel feature vector f_{kv} (defined in (10)) to the additional loss coefficients L_{ac} (given by (2)) for data-rich networks, as given by:

$$\begin{bmatrix} L_{ac,1} \\ \vdots \\ L_{ac,n_{net}} \end{bmatrix} = \begin{bmatrix} f_{kv,1} \\ \vdots \\ f_{kv,n_{net}} \end{bmatrix}\beta + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{n_{net}} \end{bmatrix} \quad (11)$$

where $L_{ac,i}$ is the additional loss coefficients for the i_{th} data-rich LV network, as defined in (2); $f_{kv,i}$ is the kernel feature vector with n_f ($n_f = 16$) columns for the i_{th} data-rich LV network; β is a coefficient vector with n_f rows; ε_i is the regression error for the i_{th} data-rich LV network; $n_{net}(n_{net} = 800)$ is the number of data-rich networks.

To obtain β and ε, an iterative algorithm is presented as follows:

1) Set $i = 0$. The ordinary linear regression [17] is used to derive coefficient vector $\beta^{(i)}$ and error vector $\varepsilon^{(i)}$.
2) According to the derived error vector $\varepsilon^{(i)}$, weighting vector w_{i+1} are given to the training samples (data-rich networks), as high weights are given to samples with low errors. This weight function is defined by:
$$w_{i+1} = \frac{1}{\varepsilon^{(i)}} \quad (12)$$
3) Set $i \to i + 1$. A weighed least square model is used to minimize:
$$\min \sum w_i \, {\varepsilon^{(i)}}^2 \quad (13)$$
After finding all w_i, $\beta^{(i)}$ is given by:
$$\beta^{(i)} = \left({f_{kv}}^T W f_{kv} \right)^{-1} {f_{kv}}^T W L_{eo} \quad (14)$$
where L_{ac} and f_{kv} are defined in (11); W is the diagonal matrix of individual weights in w_i. Correspondingly, a new $\varepsilon^{(i)}$ is derived in this step.
4) Steps 2) and 3) are repeated until the coefficient vector $\beta^{(i)}$ converges.

Detailed implementations of steps 1) – 4) are presented in [18]. After finding β, the additional loss coefficient L_{acs} for any data-scarce LV network is given by:

$$L_{acs} = f_{ksv}\beta \quad (15)$$

where L_{acs} is a scalar. f_{ksv} is the kernel feature vector of the data-scarce network. It has n_f columns. f_{ksv} is given by (10), where f_{ksv} replaces f_{kv}. β is given by (14). β is a vector with n_f rows.

2.4. Validation

In this paper, the *k*-fold cross-validation [19] is used to validate our developed approach and derive the estimation accuracy. The reasons for using *k*-fold cross-validation are: 1) the cross-validation avoids using the same data to both develop and validate the developed model; and 2) it ensures a satisfactory tradeoff between bias and variance. In each iteration of the cross-validation, a portion of the data-rich networks are reserved in the validation set and are treated as if they were data-scarce. Their APEL results are estimated by

applying our approach, which is trained using the rest of the data-rich networks. However, because the networks in the validation set are indeed data-rich networks, their accurate APEL results can be calculated. This allows for the comparison of the estimated APEL results against the accurate APEL results for validation.

The k-fold cross-validation is detailed as follows. Firstly, the additional loss coefficient L_{ac} (as defined in Section 2.1) are derived as the accurate values for the 800 data-rich networks. The rest steps are described in Fig.2.

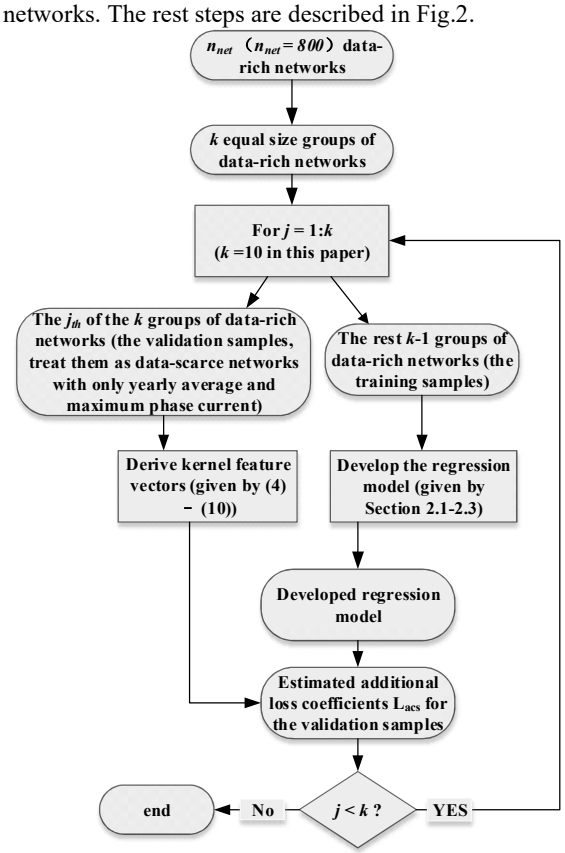

Fig. 2 *Flowchart of k-fold cross validation*

This paper uses the root-mean-square error (RMSE) to measure the regression performance. The regression performance indicates errors between the accurate values L_{ac} derived in Section 2.1 and the estimated values L_{acs} derived by applying the k-fold cross-validation to the validation samples (treat them as data-scarce networks). This error is given by:

$$e_{rmse} = \sqrt{\frac{\sum_{i}^{n_{net}} \left(L_{ac,i} - L_{acs,i}\right)^2}{n_{net}}} \qquad (16)$$

where $n_{net}(n_{net} = 800)$ is the number of validation samples. $L_{acs,i}$ is the estimated additional phase energy loss for the i_{th} validation sample (treat it as if it were a data-scarce network with only yearly average and maximum phase currents); $L_{ac,i}$ is the accurate value (derived in section 2.1) of additional phase energy loss for the i_{th}validation sample. A lower e_{rmse} indicates a better performance of the developed regression model.

2.5. Additional phase energy losses estimation for data-scarce networks

After deriving the additional loss coefficients for data-scarce networks, the APELs are estimated in two scenarios: 1) the resistances of distribution lines are available; and 2) the resistances of distribution lines are unknown.

Given a data-scarce network, its APEL is given by (3), where L_{acs} replaces L_{ac}. L_{acs} is given by (15).

For scenario 1), the APELs are directly calculated by applying (3). For scenario 2), the APELs are calculated using typical wire resistances for urban, suburban and rural networks in the UK. The typical wire resistances for urban, suburban and rural networks are 0.064Ω, 0.282Ω and 0.32Ω, respectively [20].

3. Case study

This section presents numerical results: 1) Section 3.1 gives the additional loss coefficients and corresponding features for the 800 data-rich networks; 2) Section 3.2 presents the regression results; 3) Section 3.3 presents the APEL results for data-scarce networks; and 4) a discussion is given in Section 3.4.

3.1. Data processing and feature extraction

In this section, for the 800 data-rich LV networks, the APEL are firstly derived and presented in Fig.3.

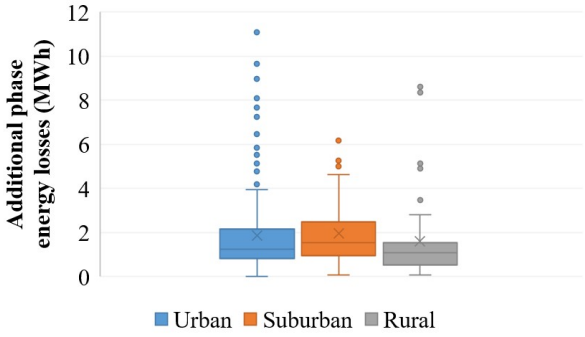

Fig. 3 *The additional phase energy losses for data-rich networks in urban, suburban and rural areas.*

Fig.3 is the range of the APELs (shown in box plot) for the 800 data-rich networks. For example, the blue dot indicates the outliers. The upper and bottom blue lines indicate the maximum and minimum APELs for urban LV networks. The line in the blue box is the average APEL for urban LV networks. The blue box indicates the range of APELs for most urban LV networks. In Fig. 3, the average APELs are 1.79 MWh, 1.95 MWh and 1.59 MWh for LV networks in urban, suburban and rural areas, respectively. For rural LV networks, the average and maximum APEL account for 0.21% and 1.21%, respectively, of the yearly distributed energy. For suburban LV networks, the average and maximum APEL account for 0.44% and 1.42%, respectively, of the yearly distributed energy. For rural LV networks, the average and maximum APEL account for 0.68% and 3.66%, respectively, of the yearly distributed energy. Furthermore, for LV networks in suburban and rural areas, the APEL account for up to: 1) 33% of the total wire energy losses; and 2) 27% of the total transformer copper losses.

Then, to develop the regression model, the additional losses coefficients L_{ac} and corresponding features (e.g.

hypothetical virtual current I_{hv}, hypothetical maximum current I_{hm}, Hypothetical degree of phase imbalance DIB_v, Root mean squares of unbalance ratio RMS) are derived. Example are given as follows:

Table 1 Examples of the additional phase energy losses coefficients and corresponding features for data-rich networks

	L_{ac}	I_{hv}	I_{hm}	DIB_v	RMS
1	1637	8.93	413.4	0.01	0.06
2	5799	53.8	614.8	0.05	0.3
3	6492	51.9	1235.3	0.02	0.11
4	2801	32.5	508.6	0.03	0.16
5	836	8.95	330.5	0.02	0.06

Thirdly, the regression error (shown in root-mean-squared error (RMSE) and mean-average-percentage error (MAPE)) from the kernel-based robust regression are used to validate the choice of these features. A lower regression error indicates a better selection of features. This validation is performed for four scenarios: 1) only I_{hv} is used as the feature to develop regression models; 2) I_{hv} and I_{hm} are used as the features to develop regression models; 3) I_{hv}, I_{hm} and DIB_v are used as the features to develop regression models; 4) excluding L_{ac}, all four features in Table 1 are used as the features to develop regression models. The validation results are presented in Table 2.

Table 2 regression error in the above scenarios

Scenario	1)	2)	3)	4)
RMSE	1163	961	715	632
MAPE	41.9%	33.4%	22.3%	19.7%

In Table 2, the RMSE and MAPE decrease with an increasing number of features used for regression. The results justify the choice of all the customised features in this paper.

3.2. Regression results

In this section, a kernel-based robust linear regression model is developed. The regression accuracy is significantly higher than that from ordinary robust linear regression. Through k-folds validation (defined in Section 2.4), the validation results are shown in Fig. 4 and Fig. 5.

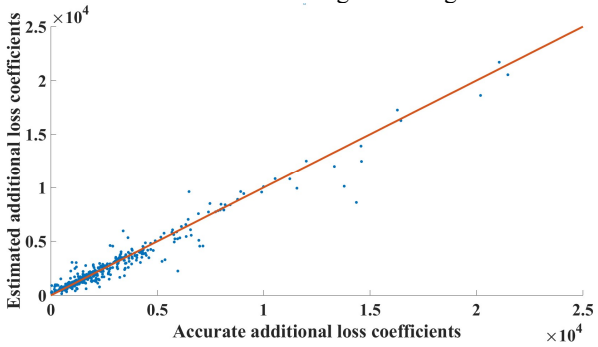

Fig. 4 *The validation results of kernel-based robust linear regression*

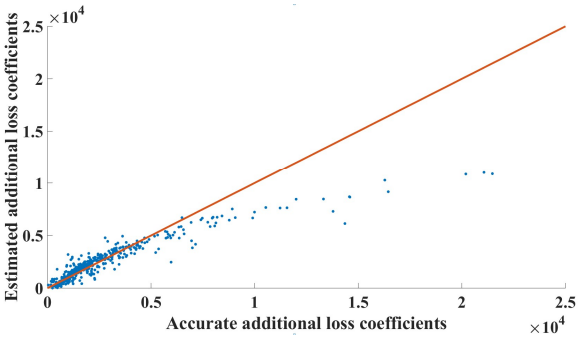

Fig. 5 *The validation results of ordinary robust linear regression*

In Fig. 4 and Fig. 5, the x-axis represents the accurate additional loss coefficients which are given by Equation (2) for 800 data-rich LV networks. The y-axis represents the estimated additional loss coefficients, when these data-rich LV networks are treated as data-scarce in the *k*-folds validation (shown in Section 2.4). The red line indicates if the additional loss coefficients are perfectly estimated by regression models. If the blue dots are closer to the red line, it indicates a higher regression accuracy. The estimated additional loss coefficients delivered by kernel-based robust linear regression are much closer to the red line than that from ordinary robust linear regression. The root-mean-squared error (RMSE) delivered by kernel-based robust linear regression is 632, which is 43% lower than that by ordinary robust linear regression.

Furthermore, the kernel-based robust regression achieves a higher regression accuracy compared to other classic regression methods. The comparison is given as follows:

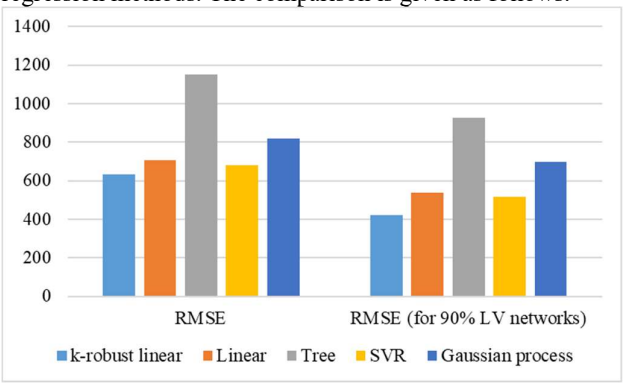

Fig. 6 *Comparison of the regression methods*

In Fig. 6, the kernel-based robust linear regression achieves almost the same RMSE as that by the support vector machine. However, when excluding 10% outliers (which presents lower regression accuracies than most LV networks), the RMSE, delivered by kernel-based robust linear regression, is lower than that from the support vector machine by 12% and other regression methods by up to 37%. Our methodology has an RMSE of slightly above 400, whereas alternative methods have RMSE values of above 500. The reduction in RMSE is attributed to the robust linear regression, kernel transformation and the customisation of features in our methodology. Further, when excluding 10% outliers, the k-robust linear regression only incurs a MAPE of 13%, i.e. on average, the estimated APEL is only 13% away from its accurate value. This estimation error is acceptable as these

data-scarce LV networks only have the yearly average and maximum phase currents. However, linear regression, tree regression, SVR and Gaussian process regression incur greater MAPEs of 17.3%, 32.7%, 16.5% and 23.9%, respectively.

3.3. *Assessments of additional phase energy losses for data-scarce networks*

After developing the regression model and calculating the additional loss coefficients for data-scarce LV networks, the APELs are derived by (3), where L_{acs} replaces L_{ac}. L_{acs} is given by (15). The k-folds validation results are shown as follows:

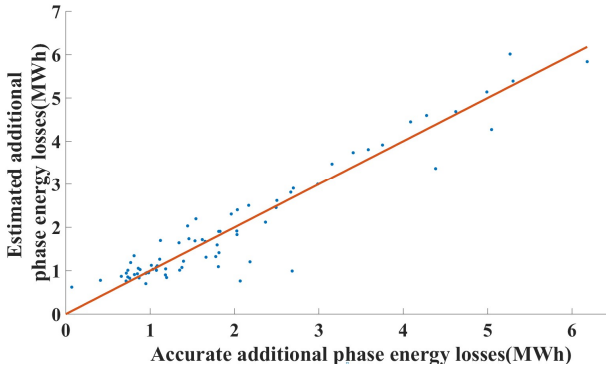

Fig. 7 *The estimation of additional phase energy losses for LV networks in urban areas*

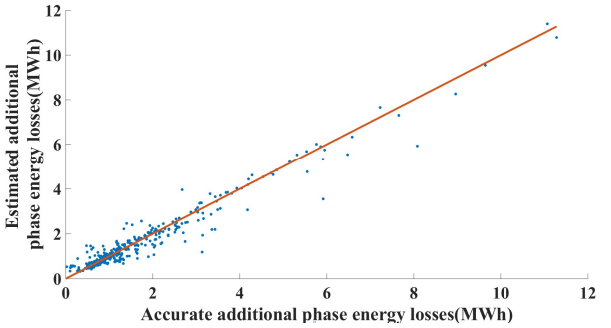

Fig. 8 *The estimation of additional phase energy losses for LV networks in suburban areas*

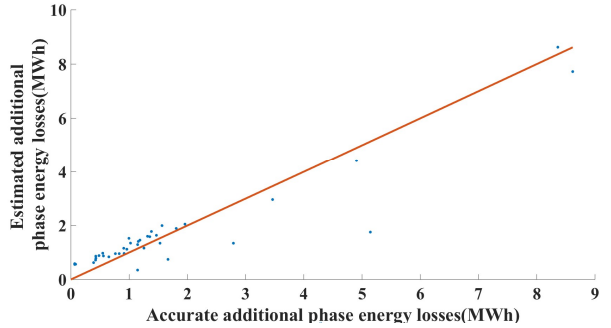

Fig. 9 *The estimation of additional phase energy losses for LV networks in rural areas*

In Fig. 7, the estimated average APEL are 1.746 MWh (which costs £314 if the electricity price is £0.18/kWh) for data-scarce urban LV networks. The average estimation error is 19.14% for 90% of the urban networks. In Fig. 8, the estimated average APEL are 1.954 MWh (which costs £352 if the electricity price is £0.18/kWh) for data-scarce suburban LV networks. The average estimation error is 11.81% for 90% of the suburban networks. In Fig. 9, the estimated average APEL are 1.531MWh (which costs £276 if the electricity price is £0.18/kWh) for data-scarce rural LV networks. The average estimation error is 12.19% for 90% of the data-scarce LV networks in rural areas.

The following figure presents the estimation accuracy of the proposed approach for LV networks with different imbalance degrees, which are defined in [21].

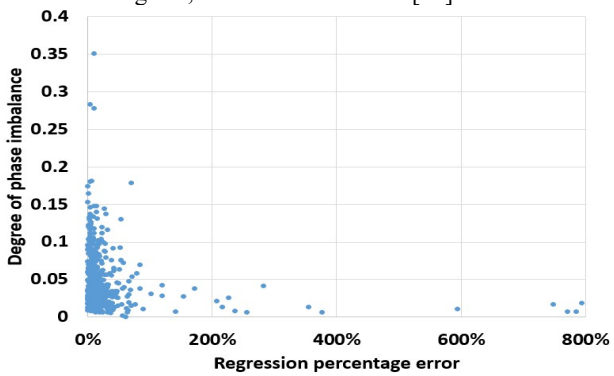

Fig. 10 *Regression errors for LV networks with different degrees of imbalance*

In Fig. 10, with the increase of the degree of phase imbalance, our proposed method delivers lower percentage error, i.e. a higher estimation accuracy is achieved for highly imbalance LV networks. For LV networks with 0.1 or higher degrees of phase imbalance, the average percentage regression error is 11.7%.

3.4. *Discussions*

In this study, our developed approach delivers about 13% percentage error in estimating the APEL for 90% of the data-scarce LV networks. This error is satisfactory because the developed approach uses minimal data (e.g. yearly average and maximum phase currents, which exists in most LV networks) to assess the year-round APEL for data-scarce networks. A higher regression accuracy can be derived if more input data are used for data-scarce networks. A trade-off is thus required by the DNOs, i.e. the DNOs should decide if it is worth to collect more data for a slightly higher regression accuracy, as more input data means more costs on data collection. In addition, for LV networks in urban area, the estimation error of APEL is higher than that for LV networks in suburban and rural areas by up to 50%. However, the higher estimation error for urban networks is acceptable. It is because according to this study urban networks correspond to very minimal APEL (only £165.6 which accounts for 9.5% of the APEL for rural networks), which are not the focus for the DNOs. For the critical focus networks (e.g. LV network which presents higher APEL in suburban and rural areas), this study delivers significant lower estimation errors, which are 11.81% and 12.19% for suburban and rural networks, respectively.

To apply this method in other countries, two points should be considered when choosing the data-rich networks: 1) there should be at least 800 data-rich networks to be collected; and 2) these data-rich LV networks should be representative.

They should cover a good mix of geographical areas (urban, suburban, and rural) and customer composition (domestic, commercial, and industrial). A higher estimation accuracy would be achieved if the training data are more representative.

For the DNOs, this paper developed an effective and efficient approach to assess the APEL. For 90% of the data-scarce LV networks (excludes 10% outlier networks), the estimation error is about 13%. In this study, it is appropriate to exclude these 10% outliers. It is because all these outliers have low APEL.

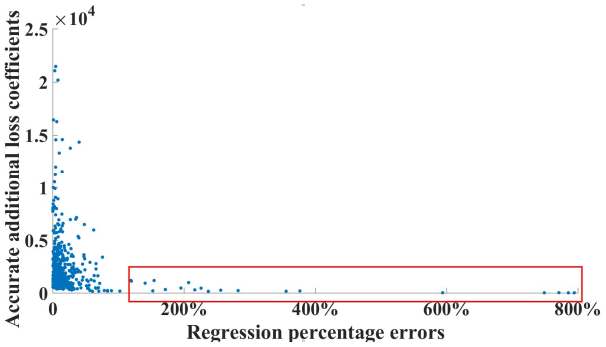

Fig. 11 *Regression errors for outlier networks*

In Fig. 11, the outliers are enclosed in the red box. They present significant regression errors by up to 800%. However, these outliers show very low APEL, which are only up to 0.3MWh ((which costs £54 additional losses if the electricity price is £0.18/kWh)). These outliers are thus out of focuses by the DNOs. Furthermore, for LV networks with significantly higher APEL, this study delivers much lower estimation errors. It is therefore appropriate to exclude these outliers.

For data-scarce LV networks, the available data of maximum phase currents can be directly obtained from maximum phase current indicators. The yearly average phase currents can be obtained through: 1) the remote telemetry unit (RTU) device on the high voltage side of LV transformers. The data on high voltage side are then transformed referred to the low voltage side. 2) The relay protection device if the device has metering function [22]. 3) The energy meters if they record the data of the three phases separately. In addition, a recent project, OpenLV, sponsored by Western Power Distribution and undertaken by EA Technology, monitors a range of LV (11kV/415V) substations and the collected data include the average phase current values [23].

It is appropriate to use regression methods for assessing the additional phase energy losses for data-scarce LV networks. This is because: 1) it is common to use regression methods to estimate or predict unknowns in both data science [17], [15] and power systems [24], [25]. 2) Through k-fold cross-validation, our approach delivers a satisfactory regression accuracy, where the average percentage error is 13% for 90% of the LV networks.

For LV networks which have high APELs (over 2.5 MWh) throughout a year, the approach delivers an accuracy of 87.3%, which is greater than the accuracy of the methodology from reference [9] by 23.7%. For LV networks less than 2.5 MWh APELs, this paper and reference [9] deliver similar estimation accuracies.

For comparison, the additional energy losses are also calculated by applying power flow analysis. However, the power flow analysis incurs unacceptably large errors when estimating the APELs for data-scarce LV networks. Given any data-scarce LV network with only yearly average phase currents and no topology, the process for calculating APEL through power flow analysis is detailed as follows: 1) assuming the loads are distributed in a rectangle distribution [13], calculate the energy losses using the unbalanced yearly average phase currents as the input. 2) Calculate the energy losses using the balanced yearly average phase currents as the input. 3) Calculate the APEL, which is the difference between the energy losses obtained in Steps 2) and 3). Through validation, when excluding 10% outliers, the power flow analysis incurs an average MAPE of 237% in the estimation of the APELs for the 800 LV networks. This error is unacceptably large, proving that the power flow analysis is not suitable for the estimation of the APELs for data-scarce LV networks. In contrast, the methodology developed by this paper is suitable for this task and it incurs the minimum error compared to alternative methods.

4. Conclusions

This study resolves a previously unanswered question: to assess the additional phase energy losses caused by phase imbalance for data-scarce low voltage (415V, LV) networks. To this end, a new statistical approach is developed with customised features. The approach learns the knowledge from 800 data-rich LV networks and then infers the additional phase energy losses for data-scarce LV networks.

Case studies reveal that: for 90% of the data-scarce LV networks in urban, suburban and rural areas, the average regression accuracies are 80.6%, 88.2% and 87.8%, respectively. These accuracies are satisfactory, as our developed approach uses minimal data (only yearly average and maximum phase currents) to assess the additional phase energy losses.

References

[1] "HV and LV Phase Imbalance Assessment," https://www.spenergynetworks.co.uk/userfiles/file/HVandLVPhaseImbalanceAssessment16.pdf.

[2] P. Carvalho, L. Ferreira, J. Santana, A. Dias, and J. Machado, "Combined Effects of Load Variability and Phase Imbalance Onto Simulated LV Losses," *IEEE Transactions on Power Systems,* vol. 33, no. 6, pp. 7031-7041, 2018.

[3] J. Wang, N. Zeng, and H. Hao, "Three-phase imbalance prediction: A hazard-based method," in 2016 IEEE International Conference on Power and Renewable Energy (ICPRE), 2016, pp. 226-231.

[4] ""LV network templates for a low-carbon future" "; https://www.westernpower.co.uk/docs/Innovation/Closed-projects/Network-Templates/LVNT-Appendix-A-Knowledge-Management.aspx.

[5] O. K. Ignatius, A. K. Saadu, and O. S. Emmanuel, "Analysis of copper losses due to unbalanced load in a transformer," *IJRRAS,* vol. 23, pp. 46-53, 2015.

[6] L. F. Ochoa, R. M. Ciric, A. Padilha-Feltrin, and G. P. Harrison, "Evaluation of distribution system losses due to load unbalance," in 15th Power Systems Computation Conference PSCC 2005, 2005.

[7] D. I. H. Sun, S. Abe, R. R. Shoults, M. S. Chen, P. Eichenberger, and D. Farris, "Calculation of Energy

Losses in a Distribution System," *IEEE Transactions on Power Apparatus and Systems,* vol. PAS-99, no. 4, pp. 1347-1356, 1980.
[8] T.-H. Chen, "Evaluation of line loss under load unbalance using the complex unbalance factor," *IEE Proceedings - Generation, Transmission and Distribution,* 142, https://digital-library.theiet.org/content/journals/10.1049/ip-gtd_19951708, 1995].
[9] L. Fang, K. Ma, R. Li, Z. Wang, and H. Shi, "A Statistical Approach to Estimate Imbalance-Induced Energy Losses for Data-Scarce Low Voltage Networks," *IEEE Transactions on Power Systems*, pp. 1-1, 2019.
[10] K. Malmedal, and P. K. Sen, "A Better Understanding of Load and Loss Factors," in 2008 IEEE Industry Applications Society Annual Meeting, 2008, pp. 1-6.
[11] L. Jiang, J. Meng, Z. Yin, Y. Dong, and J. Zhang, "Research on additional loss of line and transformer in low voltage distribution network under the disturbance of power quality," in 2018 International Conference on Advanced Mechatronic Systems (ICAMechS), 2018, pp. 364-369.
[12] A. L. Shenkman, "Energy loss computation by using statistical techniques," *IEEE Transactions on Power Delivery,* vol. 5, no. 1, pp. 254-258, 1990.
[13] W. H. Kersting, *Distribution System Modeling and Analysis, 4th Edition*, pp. 39-77: CRC Press, taylor& Francis Group, 2017.
[14] M. T. Bina, and A. Kashefi, "Three-phase unbalance of distribution systems: Complementary analysis and experimental case study," *International Journal of Electrical Power & Energy Systems,* vol. 33, no. 4, pp. 817-826, 2011/05/01/, 2011.
[15] C. Yu, W. Yao, and X. Bai. "Robust Linear Regression: A Review and Comparison," https://arxiv.org/pdf/1404.6274v1.pdf.
[16] S. R. Gunn. "Support Vector Machines for Classification and Regression," http://m.svms.org/tutorials/Gunn1997.pdf.
[17] G. A. F. Seber, and A. J. Lee, *Linear Regression Analysis*, 2rd edition ed., p.^pp. 2-4: Wiley Series in Probability and statistics, 2012.
[18] R. Andersen, *Modern Methods for Robust Regression* SAGE Publishing, 2007.
[19] J. D. Rodriguez, A. Perez, and J. A. Lozano, "Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 32, no. 3, pp. 569-575, 2010.
[20] Y. Zhang, "Cost reflective network pricing for high voltage and low voltage distribution networks," University of Bath, 27 Jun 2012.
[21] K. Ma, R. Li, and F. Li, "Quantification of Additional Asset Reinforcement Cost From 3-Phase Imbalance " *IEEE Transactions on Power Systems,* vol. 31, no. 4, pp. 2885 - 2891 July, 2016.
[22] "Sepam™ Series 20 Protective Relays User's Manual," https://www.schneider-electric.com/resources/sites/SCHNEIDER_ELECTRIC/content/live/FAQS/221000/FA221290/en_US/63230-216-208C1_Sepam_Series_20_User_Manual.pdf.
[23] R. Ash, T. Butler, R. Potter, and D. Hollingworth. "OPEN LV," https://openlv.net/wp-content/uploads/2017/10/OpenLV-Measurement-Points-V1.0.pdf.
[24] M. Mohanpurkar, and S. Suryanarayanan, "Regression Modeling for Accommodating Unscheduled Flows in Electric Grids," *IEEE Transactions on Power Systems,* vol. 29, no. 5, pp. 2569-2570, 2014.
[25] J. Zhang, C. Y. Chung, and Y. Han, "Online Damping Ratio Prediction Using Locally Weighted Linear Regression," *IEEE Transactions on Power Systems,* vol. 31, no. 3, pp. 1954-1962, 2016.